
id:analytics™

Department of Homeland Security

DATA MINING WORKSHOP

Stephen Coggeshall

Chief Technology Officer

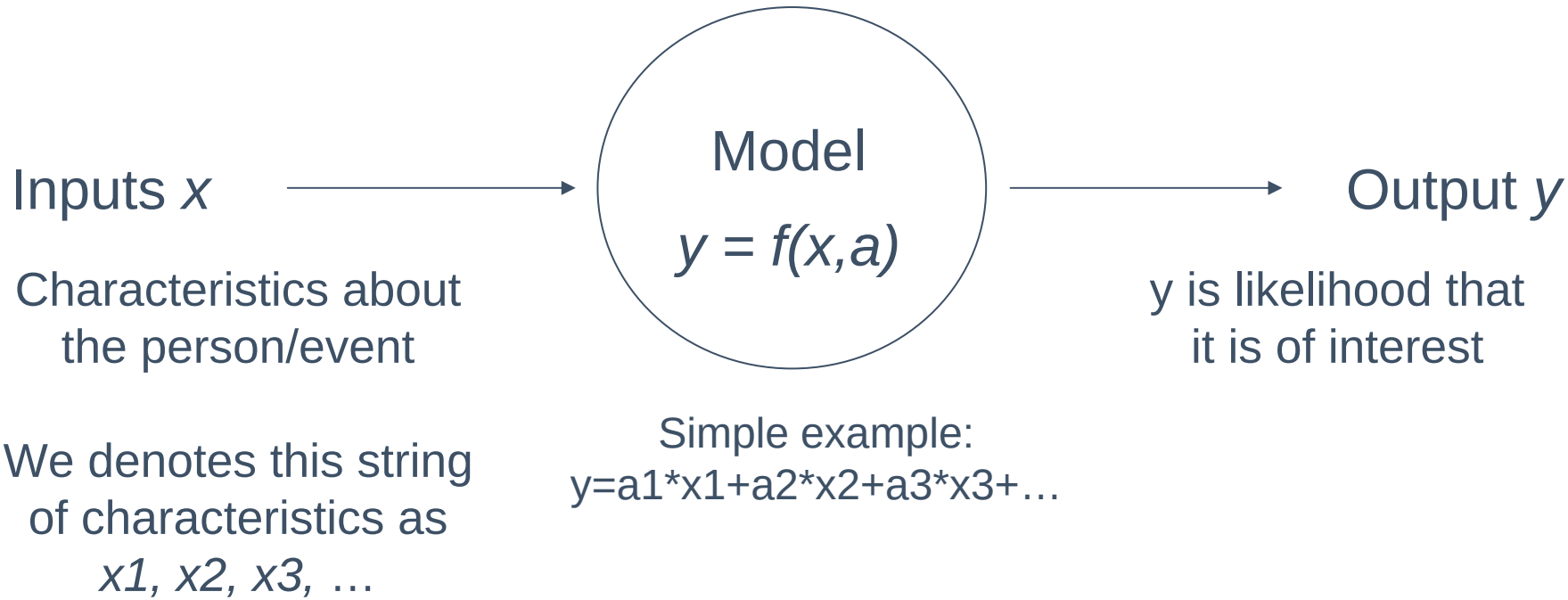
ID Analytics

July 24-25, 2008

Overview

- What is a data mining model?
- How to build a model?
- What to do when you don’t have known bads?
- How to evaluate a data mining model when you don’t have known bads?
- What’s the benefit of using models?

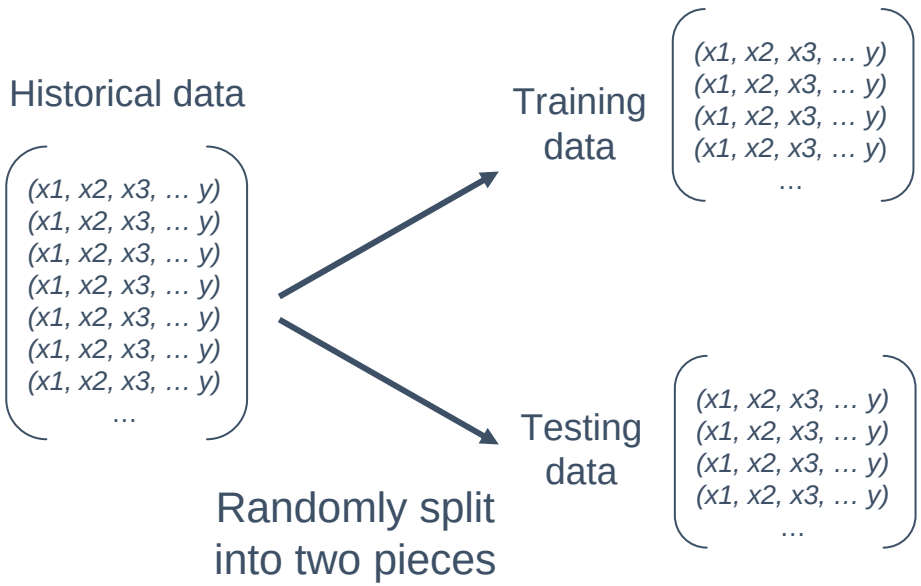
What is a Data Mining Model?



The "*a*" in the model is a set of parameters that are "learned" from data before the model is to be used

How to Build a Data Mining Model?

- We use LOTS of data to build a model (the more the better)
- A data record looks like this: (x1, x2, x3, …, y)
- We use many (millions) of data records
- Very important to clean the data as much as possible



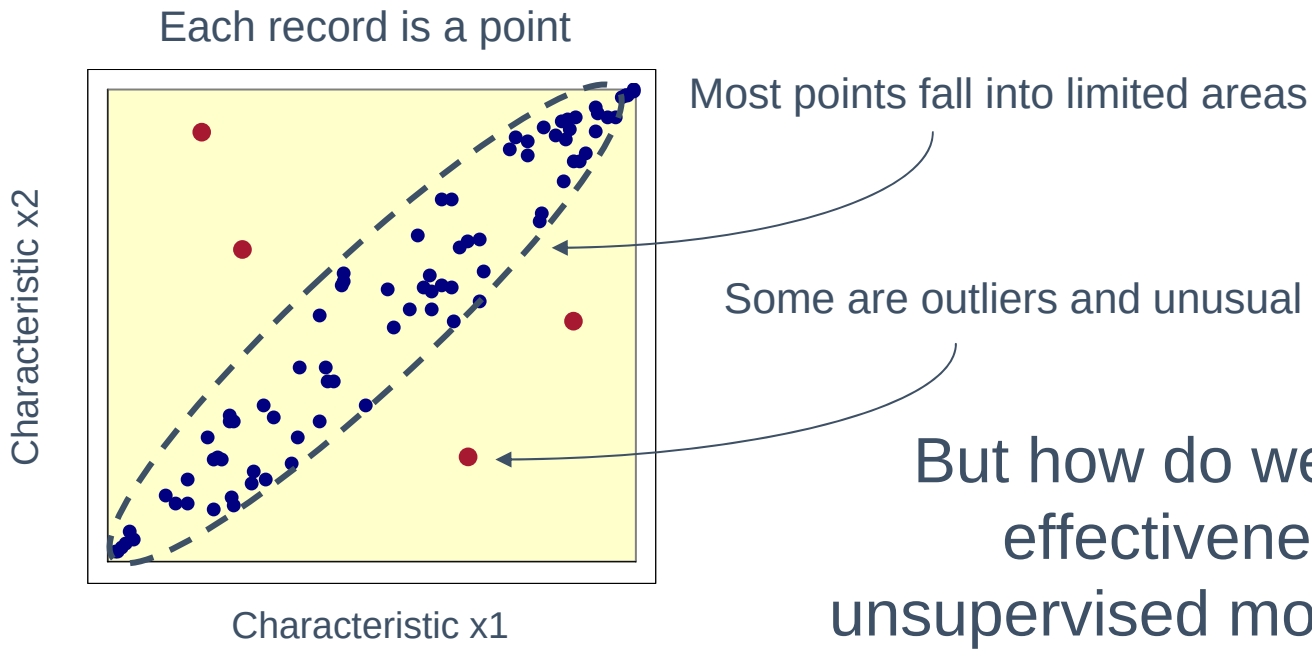
Use statistical/machine learning algorithms to train the model: find best parameters *a* in $y = f(x,a)$

Test/validate the model with data the model has never seen before. Quantify how well the model does at predicting *y*

This is the usual methodology and is called ***supervised training***

What To Do When There is No Outcome Data?

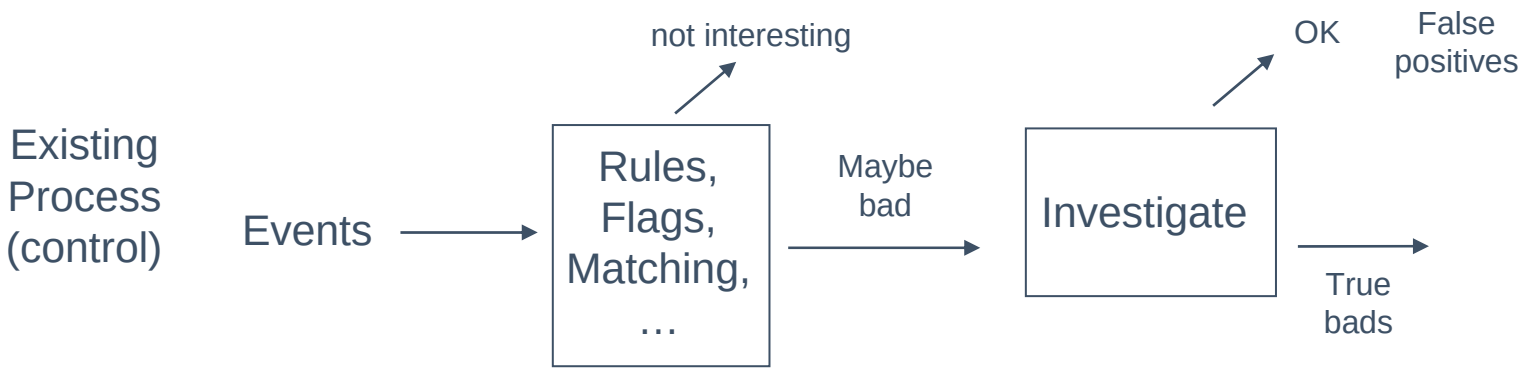
- Sometimes we don't know the outcome – who's good and who's bad
- Our data records look like this: *(x1, x2, x3, …).* We don't have a *y.*
- We can build an ***unsupervised*** model – identifies outliers/anomalies



But how do we asses the effectiveness of an unsupervised model if we have no previously known bads?

How to Validate a Model With Little or No Previously Known Bads (slide 1)

Evaluate as a test/control: First, measure efficacy of existing process



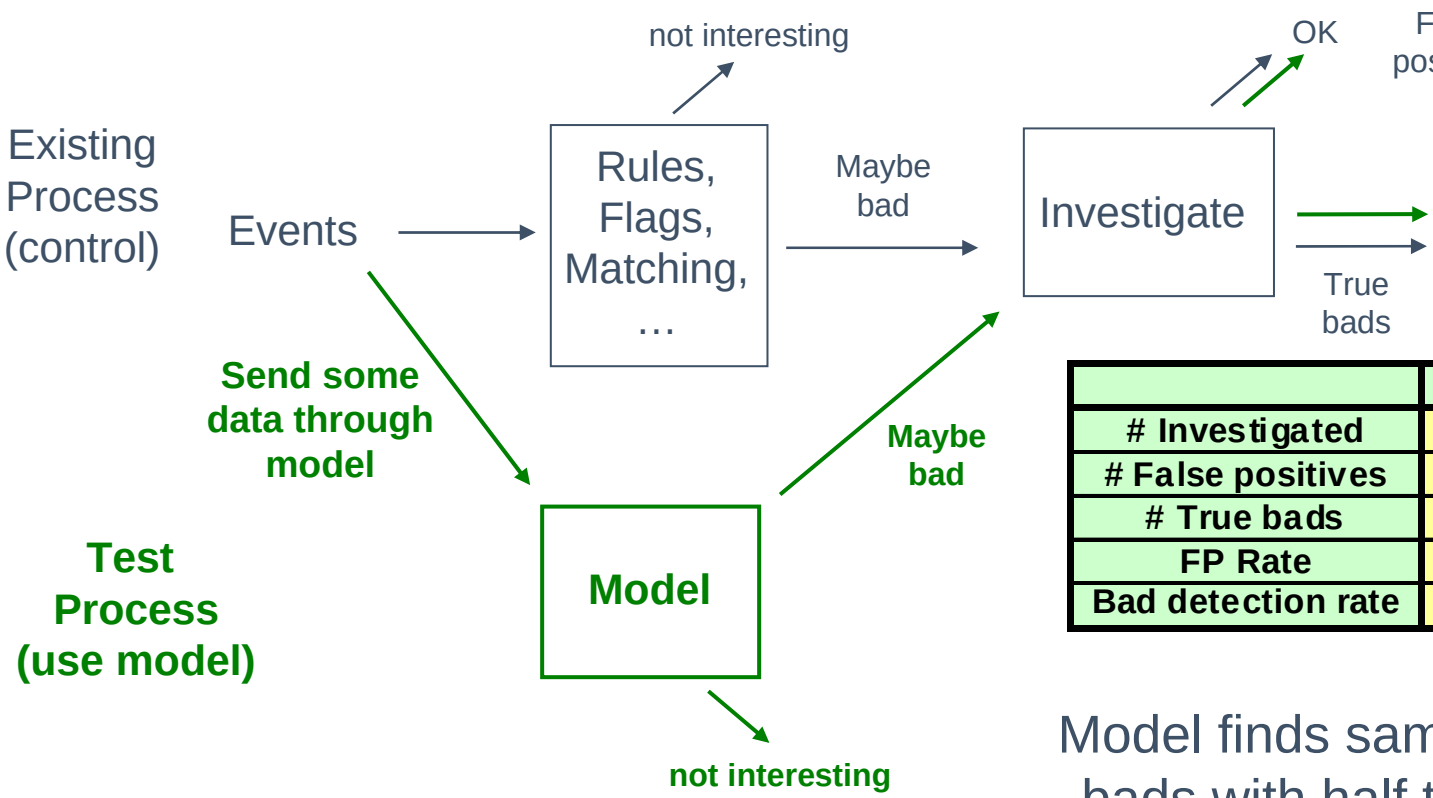
Important metrics to measure efficacy:

False positive rate = # false positives / baseline *
Bad detection rate = # true bads / # investigated

* baseline either all investigated or true bads. It's not important which one you choose.

How to Validate a Model With Little or No Previously Known Bads (slide 2)

Evaluate as a test/control: Compare model efficacy to existing process

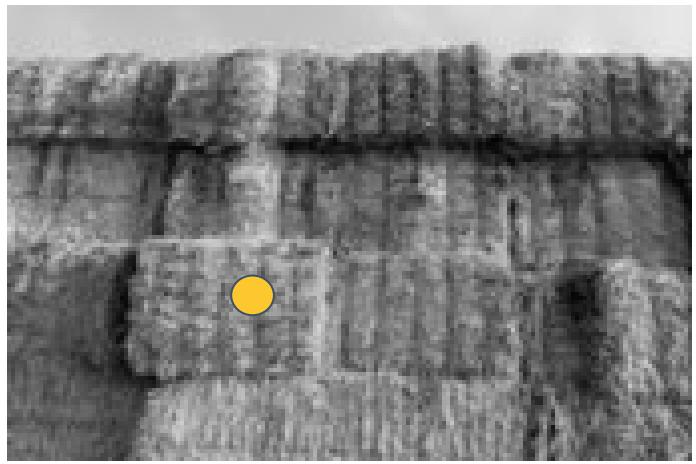


	Control	Test
# Investigated	1,000	100
# False positives	800	60
# True bads	200	40
FP Rate	80%	60%
Bad detection rate	20%	40%

Model finds same # bads with half the effort, half the intrusion

Data Mining Can Help to Find the Bads While Reducing the Size of the Examined Population

Data Mining



Human Investigation

Examples of Successful Data Mining Models in Use

Supervised Data Mining Models

- Credit risk scoring
- Credit fraud
- Identity fraud
- Bankruptcy
- Consumer products targeting
- Account attrition
- Cross sell optimization
- Segmentation
- Customer value and profitability
- Product migration
- Econometric forecasting
- Mutual funds redemption
- Corporate valuations
- Stock market trading
- Derivatives pricing
- Commodities pricing
- Bonds pricing

Unsupervised Data Mining Models

- IRS taxpayer/preparer fraud
- Healthcare claims fraud and abuse

Not an exhaustive list. These are models that my teams have built and implemented

Summary

- Data mining models work and in wide use in public and private sectors
- Build ***supervised*** models when you have known bads
- Build ***unsupervised*** models when no known bads
- Can quantitatively evaluate model effectiveness even with no previous known bads
- Data mining models can discover previously unknown relationships
- Models can minimize review population
  - Allows more efficiency and effectiveness
  - Minimizes intrusiveness

Department of Homeland Security

DATA MINING WORKSHOP

Stephen Coggeshall

Chief Technology Officer

ID Analytics

July 24-25, 2008


id:analytics™